Корпус текстов как особый лингвистический ресурс

Содержание

1 Понятие и основные характеристики корпуса текстов

1.1 Размер и репрезентативность

1.2 Разметка

Лингвистический корпус – это совокупность текстов,

- собранных в соответствии с

Пользователи корпусов

- прикладные лингвисты различного профиля;
- лексикографы;
- преподаватели (корпусы используются как база при обучении языкам);
- компьютерные лингвисты;
- другие специалисты по языку (литературоведы, редакторы, специалисты по рекламе);
- специалисты по общественным наукам (историки, социологи и др.).

Что дают корпусы пользователям

- реальные контексты;
- реальные статистические данные (на больших объемах текстов);
- сочетаемость (коллокации);
- категоризацию языкового материала;
- проекции языка на различные подъязыки.

Основные характеристики корпуса текстов

- Размер корпуса
- Репрезентативность корпуса
- Разметка

Размер корпуса

- имеет фиксированный характер;
- имеет нефиксированный характер (пополняется).

Размер корпуса

Объем – 1 млн. словоупотреблений (Брауновский корпус, Уппсальский корпус русского языка).

Первые корпусы

Объем общеязыкового (национального) корпуса должен быть не < 100 млн. словоупотреблений.

Современный подход

С течением времени объем корпуса может меняться, однако эти изменения должны или не менять ***репрезентативность*** корпуса, или менять ее обоснованно.

Репрезентативность корпуса –

это его представительность, пропорциональное соотношение его отдельных частей (по разным характеристикам – время, жанры, стили, авторы и др.).

Типы корпусов с точки зрения репрезентативности

Первый тип	Второй тип
Отражают все многообразие речевой деятельности; универсальны.	Отражают бытование какого-либо лингвистического или культурного феномена в общественной речевой практике; построены для специальной цели.

Типы корпусов с точки зрения отбора текстов

- Включаются разные по жанрам, стилям и тематике тексты, но устанавливаются пропорции, в которых должны быть представлены тексты.
- Имеют фиксированный объем.
- Пополнение происходит после тщательной процедуры отбора новых текстов.

Сбалансированные корпусы

- Баланс текстов разных стилей, жанров и тематики не соблюдается.
- Имеют большой объем.
- Постоянно пополняются новыми текстами на данном языке.

Мониторные корпусы

Разметка

Для решения различных лингвистических задач мало наличия массива текстов. Тексты должны содержать дополнительную лингвистическую и экстралингвистическую информацию – ***разметку*** (аннотацию). Разметка во многом определяет возможности, предоставляемые корпусами исследователям.

Разметка заключается в приписывании текстам и их компонентам специальных меток (tag, tags).

2 Виды разметки в корпусе

Нелингвистическая и лингвистическая разметка

Сведения об авторе и о тексте: автор, его возраст, пол, годы жизни и др.; название, год и место издания, жанр, тематика и др.

Лексические, грамматические, просодические и др. характеристики элементов текста.

Нелингвистическая разметка (метаразметка)

Лингвистическая разметка

Типы лингвистической разметки

- морфологическая
- синтаксическая
- семантическая
- просодическая
- анафорическая

К первичной лингвистической разметке текстов относятся этапы, обязательные для каждого корпуса:

токенизация (разбиение на орфографические слова);

лемматизация (приведение словоформ к словарной форме).

Морфологическая разметка – это основной тип разметки

- большинство крупных корпусов являются морфологически размеченными;
- морфологический анализ является основой для дальнейших форм анализа – синтаксического и семантического;
- успехи в компьютерной морфологии позволяют автоматически размечать корпусы больших размеров.

Морфологическая разметка включает признаки части речи и грамматических категорий, свойственных данной части речи.

Программные средства автоматического морфологического анализа — ***тэггеры*** (taggers).

Синтаксическая разметка

является результатом синтаксического анализа, выполняемого на основе данных морфологического анализа; описывает синтаксические связи между лексическими единицами и различные синтаксические конструкции.

Программные средства автоматического синтаксического анализа — ***парсеры*** (parsers).

Семантическая разметка

характеризует семантические категории, к которым относится данное слово, и более узкие подкатегории, специфицирующие его значение.

3 Основные этапы создания корпусов

1. Определение перечня источников.

2.Оцифровка текстов (преобразование в компьютерную форму).

3. Предобработка текста (филологическая корректировка, подготовка экстралингвистического описания текста).

4. Конвертирование — предварительная машинная обработка (удаление или преобразование нетекстовых элементов, удаление из текста переносов, обеспечение единообразного написания тире и др.).

5. Разметка текста. Метаразметка и собственно лингвистическая разметка.

6. Корректировка результатов автоматической разметки: исправление ошибок и снятие неоднозначности.

7. Конвертирование размеченных текстов в структуру специализированной информационно-поисковой системы (corpus manager).

8. Обеспечение доступа к корпусу: в пределах дисплейного класса, на CD-ROM или в режиме глобальной сети. Различным категориям пользователей могут предоставляться разные права и возможности.